专题专利预警平台建设方案研究与实践*

王　丽[1,2]　丁迎杰[1]　吴　鸣[1]
[1](中国科学院文献情报中心　北京　100190)
[2](中国科学院大学　北京　100049)

摘要:【目的】制定专题专利预警平台建设方案，为长期的专题跟踪预警分析、专题数据再利用等工作提供一种解决途径。【方法】平台集成开源代码平台和工具(DSpace、OpenRefine、ECharts、VOSviewer 等)，实现对专题数据的存储、跟踪、分类、清洗、分析、管理等功能。【结果】选择极紫外光刻技术专题进行应用实践，测试并解决实践过程中的细节问题，验证专题专利预警平台的可行性和有效性。【局限】目前的专题专利预警平台数据处理全自动化、数据分析指标化、内容挖掘的关联实现等方面需要进一步优化。【结论】专题专利预警平台所实现的功能，对于在技术研发生命周期内进行技术专利及时跟踪预警并分类管理有着现实的意义。

关键词: 专题管理　专利预警　平台建设
分类号: TP392　G353

1　引　言

专利信息对经济社会发展和企业创新活动有着重要的支撑作用。专利预警通过检索和分析专利信息，对可能面临的专利风险进行研究和预测，从而支撑应对策略[1]，专利预警的相关利益主体大至区域国家小到研发团队。在技术创新和产业发展生命周期中，利益主体需要及时敏锐地捕捉相关专利信息，分析和预测专利风险并做出风险应对。

由于贯穿于技术创新和产业发展生命周期专利风险是动态变化的，利益主体针对特定创新专题进行专利预警工作是有时效性的。从专利检索到专利分析，从风险捕获到风险评估，人工专利预警工作无法实时监测专利信息，过程专利数据往往仅实现一次性利用，这些弊端阻碍了专利预警的时效性及专利数据的再利用。专利数据作为专利信息的主要载体，在专利预警工作中的有效利用至关重要。在数字化的今天，网络日益成为科技交流和传播最重要渠道，利用网络平台及时跟踪专利信息实现专利预警有着现实的意义。

笔者在使用和调研国内外相关系统平台中发现:

(1) 专利分析平台，可较好地满足单次专利分析需求，但对于专利预警来说功能较为分散，有些平台可实现部分预警分析的功能，但是存在专题创建不灵活、定制化程度不高、或者无法实时跟踪等不同功能的缺失。如，Thomson Innovation 平台可以通过创建预警实现某个专题的数据跟踪但无法进行数据清洗、分类管理等。Orbit 分析平台可以通过创建工作文件夹对固定专利数据集进行数据清洗、分析，但无法实时跟踪、分类管理。

(2) 企业竞争情报系统，基本遵循情报搜集、情报分析、情报服务的逻辑框架，数据源庞杂，非结构化信息占据主力，但专题专利预警的针对性不足，并不重视多来源数据的清洗。如，谷尼竞争情报系统对各种信息源进行全面整合和利用，综合对比提供企业竞争情报资讯，但是对于技术研发创新来说信息过于庞杂。

(3) 自主建设平台，根据专利管理体系、专利情报分析体系、专利信息价值体系等自主建设相关平台，如，中国科学院计算技术研究所开发的专利价值分析

通讯作者: 王丽，ORCID: 0000-0002-9513-6159，E-mail: wangli@mail.las.ac.cn。

*本文系中国科学院文献情报中心青年项目“高技术产业的专利池知识产权共享模式研究”(项目编号: Q130071001)和中国科学院文献情报能力建设专项“中国科学院机构知识库功能扩展”(项目编号: Y5ZG081001)的研究成果之一。

与评级电子系统从专利价值评估、发明人自评、专家评审等角度进行设计与开发。

针对上述问题，本文提出一个专题专利预警平台的实现方案，该平台根据利益主体对技术创新及管理的具体需求建立，并有效结合专利情报分析流程，可实现技术或产业链各环节的定制化的专利分类管理和预警，数据主体既可以是整个行业的专利数据集合，也可以是某一具体产品或技术的专利数据集合。在功能上，利益主体利用该平台可以形成专题专利数据集、分类管理研发关键技术、监测最新技术发展动向、了解竞争对手的技术水平等，实现专利预警分析及专利数据的再利用。在技术上，基于该方案集成开源代码平台和工具并开发实现专题专利数据的存储、跟踪、分类、清洗、分析、管理等功能模块，节约了系统开发的时间和经济成本。

2　专题专利预警平台建设方案设计及实现

专利数据作为专利信息的主要载体，也是专利预警工作的主要分析对象。专利数据的开放性为本专利预警平台的采集、整理、加工提供了实现基础。专利预警需要全面及时跟踪专利信息，专利数据的开放资源众多，不同资源的元数据及内容格式并不统一。专题专利预警平台需要通过对研发专题的定制实现专题数据的自动采集，从不同的专利信息资源中，匹配制定内容且结构化抽取专利元数据信息，同时实现本地存储。专利预警平台需要通过统一定制的元数据实现不同来源专利数据的归一化并可通过设置专利数据的唯一项进行数据去重，具备数据处理功能实现内容格式的统一性，从而帮助利益主体实现专利信息的动态捕捉、跟踪，并为预警分析提供良好的数据质量。专题专利预警平台的功能框架如图 1 所示，基于这一框架重点完成预警平台的 5 项功能：专题定制、数据采集、自动分类、数据处理、数据分析。

(1) 专题定制是指对需要预警的目标进行定制，通过制定完善的检索策略实现专题定制。在制定检索策略之前需要进行充分的准备工作，如专题领域调研、主题词多样性调研、目标专利资源预检索、检索策略调整修正等，检索策略的完备性直接影响后续跟踪数据的有效性。利益主体通过专题定制实现所需专题的个性化跟踪预警。

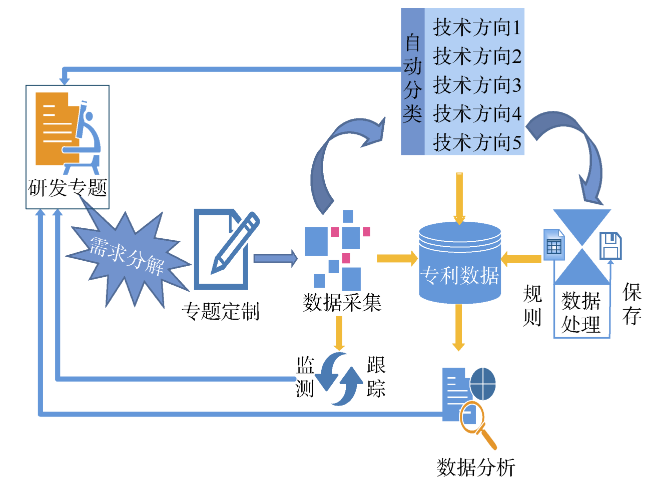


图 1　专题专利预警平台的功能框架

(2) 数据采集是指定期对目标专利资源进行采集，构建一系列可分布部署的网络定向采集器实现对目标专利资源的精准采集[2]。数据采集阶段将目标专利资源的元数据项进行归一化处理，实现不同来源数据的描述统一化，并通过专利数据唯一项进行数据去重。利益主体通过数据采集实现定制研发专题的实时跟踪。

(3) 自动分类是指对采集到的信息进行定制化分类处理。通过制定完善的匹配策略实现主题分类管理，对于每一条采集到的专利信息进行主题识别，并自动分配到相应的主题分类里，同时实现专利信息的自动标引。利益主体通过自动分类实现定制专题的分类跟踪及管理。

(4) 数据处理是指对采集到的专利信息进行数据处理，包括数据清洗、格式处理等。专利信息资源的多元化导致著录格式不统一，数据采集阶段对元数据进行归一化命名处理，但采集到的专利信息著录规则存在多样性，数据处理阶段的目的之一是将著录格式统一化。未经清洗的数据普遍存在命名不规范，如 IBM 公司可能存在 IBM、IBM Corp.、International Business Machines Corporation 各种形式的写法，且存在下属公司、各地区分公司以及其他法人机构，如不通过数据清洗加以规范，针对申请人的跟踪预警分析则失去准确性，因此数据处理阶段的另一个目的是实现数据清洗。利益主体通过数据处理规则的保存实现预警信息的规范有效性。

(5) 数据分析提供面向利益主体的自动预警分析服务。通过上述一系列工作，数据分析阶段呈现包括重点机构的跟踪、重要发明人的揭示、热点主题的揭

示等服务，同时利益主体可根据需求实现多角度的分析，深度发掘预警信息。

3 专题专利预警平台的关键技术实现方法

根据专题专利预警平台的功能框架，基于开源软件 DSpace 4.2[3-4]进行扩展开发，并集成开源软件 OpenRefine、ECharts 及 VOSviewer 实现相关的功能。DSpace 是基于 Java 的开源系统，其具有完善的元数据定义、数据的本地化分层存取、数据的索引与检索，成为专题专利平台开发建设的首选系统。基于 DSpace 的元数据功能，专题专利预警平台对监测的目标数据字段进行统一规范。DSpace 的社群(Community)和集合(Collection)为专题专利预警平台的分类管理提供了技术实现基础。为了实现更加准确的预警分析，专题专利预警平台需要对采集后的数据进行规范化处理，OpenRefine[5]的数据处理功能及开源性成为实现数据处理功能的首选，平台集成 OpenRefine 实现对采集数据的清洗、处理以及处理规则的保存。DSpace 系统提供一维的统计分析，平台通过二次开发实现自定义多维组合分析，从而拓宽预警深度及广度，并集成 ECharts 及 VOSviewer 实现分析可视化。专题专利预警平台的技术框架如图 2 所示：

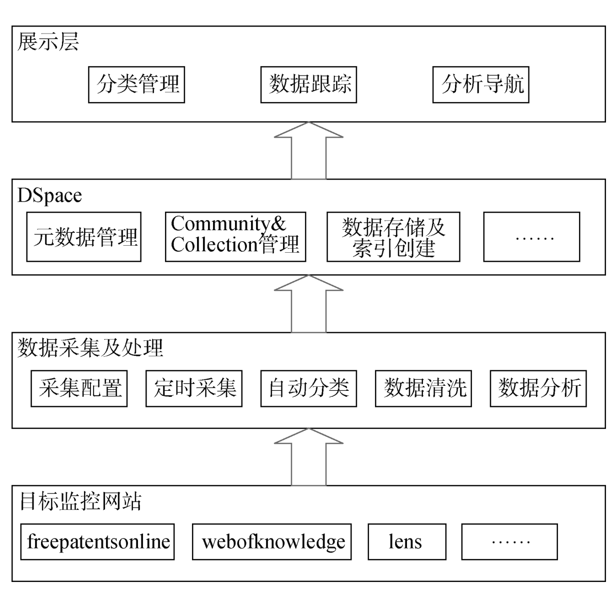


图 2 专题专利预警平台的技术框架

专题专利预警平台的建设过程中要解决的关键点是数据的定制化采集、自动分类、数据清洗、数据分析 4 个关键技术。关键技术实现流程如图 3 所示：

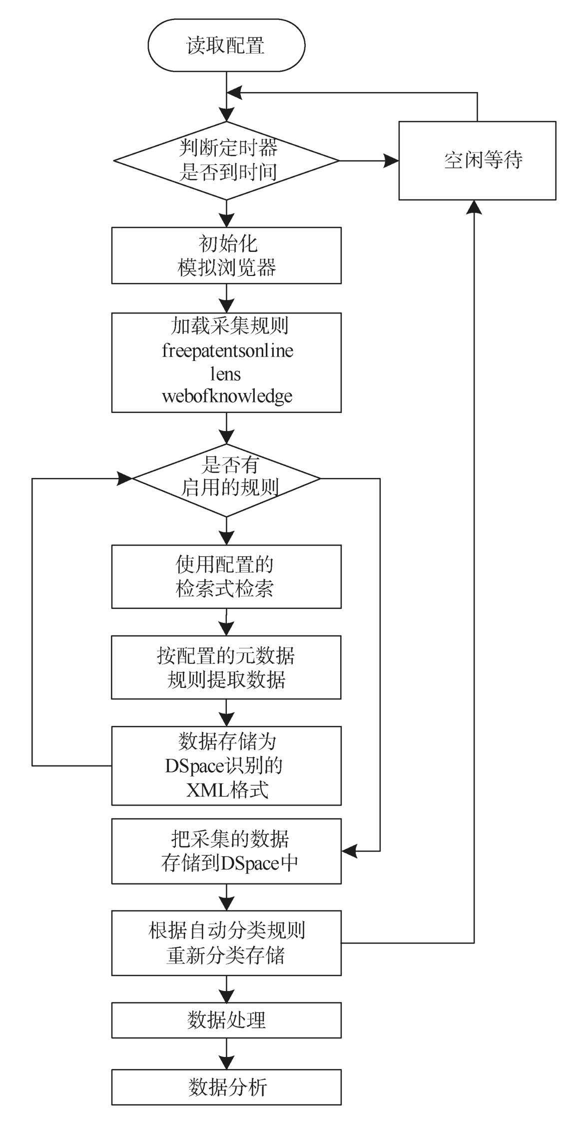


图 3 专题专利预警平台关键技术实现流程

3.1 数据采集

为了实现全面的监控，目标专利数据可能来源于不同的专利信息网站，网页的内容结构差别较大，同时网站的更新升级也会导致网页结构的变化，这对系统的适应性提出了较高的要求。专利预警平台需要具备灵活的分析能力，针对目标专利信息资源的网页内容进行自动分析和采集，将采集的数据按照事先定义好的元数据进行本地化存储。系统主要采用 Apache 的 HttpClient 模拟浏览器的功能：

(1) 采用单例模式封装 HttpClient，并利用 Double Check 解决在多线程采集时创建多个实例的问题，如图 3 所示的初始化模拟浏览器。

(2) 利用 HttpClient 模拟浏览器，实现采集功能。

不同的专利信息网站的检索规则有所差别，将事先定制好的检索策略按照目标专利信息资源网站的检索规则进行检索式配置，在数据采集过程中将根据配置的检索式进行数据检索，实现过程与人工检索相似。

(3) 对采集的 HTML 格式数据采用 jsoup 组件进行提取，通过配置 HTML 元素选择器与 DSpace 的元数据形成一一对应关系，从而把专利信息内容按配置好的元数据格式存储到本地。

3.2　自动分类

为了实现技术或产业链各环节或行业各分支的分类管理和预警，专利预警平台的自动分类预警功能是必要的，本研究中采用分类配置实现自动分类管理。分类配置信息采用 XML 的方式进行存储，并与 DSpace 中的社群(Community)和集合(Collection)进行关联，从而实现数据的分类存储。首先通过社群(Community)实现技术或产业链各环节或行业各分支的预定分类设置，便于专利预警平台的前端展示。其次通过集合(Collection)进行存取分类定义，即对每个存取分类设置匹配规则，从而使采集到的数据根据定义规则进行分类存取从而实现专利数据的自动分类。自动分类的实现利用了 DSpace 强大的搜索功能，对于符合预设分类规则的数据，利用 DSpace 的内置应用程序接口，移动到相应的集合(Collection)里去。

3.3　数据处理

尽管数据采集阶段进行了元数据统一化处理，但是各信息网站的数据仍存在著录格式不统一、内容表达不统一等问题，因此专题专利预警平台的数据处理及清洗功能必不可少。数据处理开源工具 OpenRefine[6]迎合了需求，但是 OpenRefine 的数据清洗是不可以重复的，一次的数据处理过程不能作为模板进行下一次数据处理，需要对其进行二次开发，实现数据清洗模板化，以便对更新的监测数据自动进行相同数据处理，从而避免了重复工作。

OpenRefine 是基于项目的数据处理，内部的处理记录是针对数据的变化，并没有对操作步骤进行记录，因此不能作为通用的数据处理规则。经分析发现 OpenRefine 的数据处理是基于命令的模式进行的，也就是每个操作都是一个命令，这样只需要对每个命令操作进行持久化存储，就可以记录 OpenRefine 的操作步骤，然后再针对存储的操作步骤，针对不同的数据进行回放操作，即实现了利用 OpenRefine 进行数据处理的模板化。

3.4　数据分析

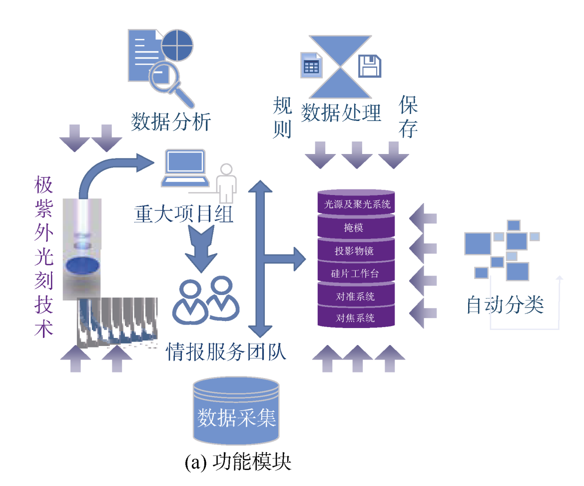
经过采集、分类、清洗等过程专利数据已优质地分类存储在专利预警平台中，实现专利数据的及时跟踪。专利预警平台需要数据分析来深度发掘专题专利信息达到多维预警效果。DSpace 系统可实现一维专利数据分析，如对机构、发明人、时间等一维统计分析，为了增强专利预警平台的数据分析功能，使其可以进行多维组合分析，本研究进行了二次开发。

专题专利预警平台利用 Solr 的 Facet 功能作为数据分析的开发基础。一维统计分析可直接利用 Solr 的 Facet 功能，系统开发元数据显示选择功能，从而实现前端界面定制化显示。实现定制化组合分析[5]，首先需要确定分析维度。其次在系统实现上，针对第一维度进行统计分析，然后基于统计出的数据子集与第二维度进行组合分析。例如进行主要机构的趋势分析时需要确定分析组合是机构和年份，分析时需要对机构进行 TOP 查询，取出 TOP 机构的子数据集，再针对该子数据集利用 Facet 的统计功能，分析其年度数据(可以指定时间区间或离散值)，最后利用系统集成的 ECharts[7]进行组合分析可视化。专题专利预警平台对内容的文本挖掘进行了初探，集成 VOSviewer[8]进行主题聚类分析，但目前的聚类与数据的关联性不足，也是下一步工作待解决的问题。

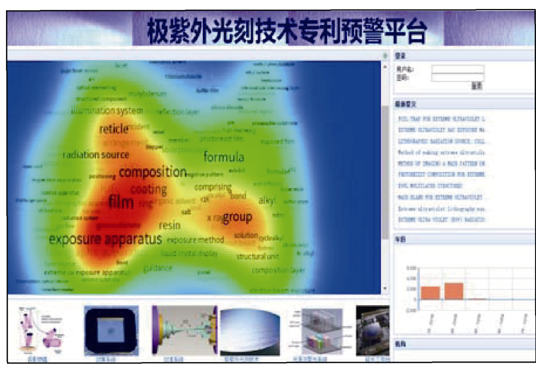
4　专题专利预警平台的应用实践

以极紫外光刻技术专利预警平台建设实践为例，对本文提出的专利预警分析平台的实现加以说明，极紫外光刻技术专利预警平台可实现极紫外光刻技术专利数据的自动采集、实时更新、数据清洗、数据分析，并且结合极紫外光刻技术的实际需求，实现了自动分类功能，如图 4 所示。

专题专利预警平台根据专利数据特征制定了统一的元数据规则(适用于不同来源的所有专利数据)，如图 5 中名称、元数据 element、元数据 qualifier 等对于所有专利数据都是统一的。不同来源的数据需要配置针对性的 HTML 元素选择器，从而将其与本平台的元数据对应。如图 5 所示，极紫外光刻技术专利预警平台针对来源于 freepatentsonline 的专利数据进行元数



(a) 功能模块



(b) 首页

图 4　极紫外光刻技术专利预警平台

据配置，从而实现平台数据结构的统一化。平台通过专利申请号即“Application Number”作为唯一标识进行数据去重处理，在此阶段原始待采集数据 7 356 条，经过系统去重后的数据为 5 787 条。平台实时监测更新数据在首页最新提交模块呈现，实现预警跟踪。

元数据

保存　添加　删除

索引	名称	元数据elemen	元数据qualif	数据类型	匹配值	数据类型	选择器	值表达式
0	Detail URL	doc	url	搜索列表	3	detail	a	@href
1	Title	title		明细	Title:			
2	Abstract	description	abstract	明细	Abstract:			
3	Inventors	contributor	Inventor	明细	Inventors:			
4	Application Number	identifier	ApplicationN	明细	Application Number:			
5	Publication Date	date	Publication	明细	Publication Date:			
6	patent Image	doc	attachmentur	明细	View Patent Images:	link	a	@href
7	doctype	doc	type	固定值				patent
8	Assignee	contributor	Assignee	明细	Assignee:			
9	Application Date	date	Application	明细	Filing Date:			
10	Application Year	date	Applicationy	明细	Application Year:			
11	Primary UPC	subject	PrimaryUPC	明细	Primary Class:			
12	UPC	subject	UPC	明细	Field of Search:			
13	IPC	subject	IPC	明细	International Classes:			
14	US Patent Reference:	relation	Reference	明细	US Patent References:			
15	Patent infor	identifier	Patent	明细			label.float_	

图 5　极紫外光刻技术专利预警平台的元数据配置

根据极紫外光刻预警平台的分类跟踪、管理需求(见图 4(a))，对各分类配置相应规则，如图 6 所示。从而实现专利数据的自动分类，对于规则无法识别的数据可通过人工判读进行补充。极紫外光刻技术的分类管理可在首页实现导航(见图 4(b))。

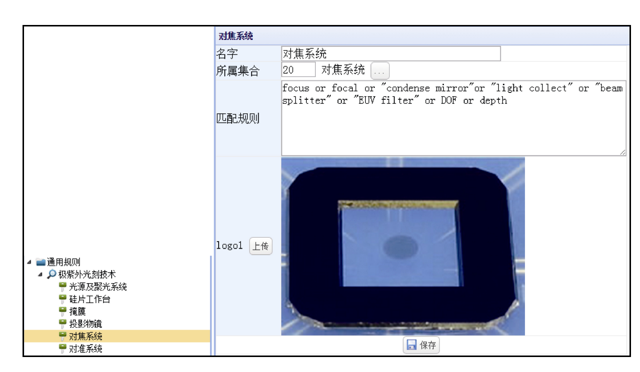


图 6　极紫外光刻技术专利预警平台的自动分类配置

良好的数据是得到精确专利预警分析结果的基础，完成极紫外光刻技术专利数据采集之后，通过本平台的数据处理功能进行数据清洗，得到统一的、规范的数据。如图 7 所示，未经处理的极紫外光刻技术专利数据中，尼康公司拥有的专利数量为 148 件，经过处理后其数量变为 266 件，可见未经处理的数据会误导专利预警结果。本专利预警平台的数据处理过程会自动保存，如图 7 所示，以便后续实时更新数据的再处理。

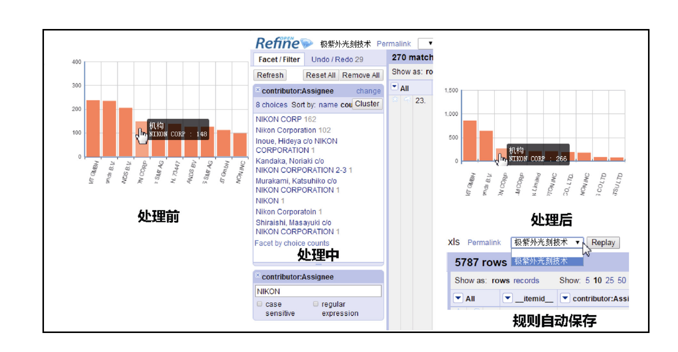


图 7　极紫外光刻技术专利预警平台的数据处理过程

经过数据处理规则的制定，极紫外光刻技术专利预警分析平台已经可以很好地实现自动采集、分类管理、实时更新，并具备良好的预警分析数据。通过平台的分析功能实现极紫外光刻技术的专利预警分析，且分析结果可以灵活地定制在平台首页展示，实现导航。极紫外光刻技术专利预警平台首页提供了专利申请时间、专利申请人、专利发明人等的预警分析导航模块。本平台可实现二维组合分析，用户可根据具体需求自由组合进行二维分析并可视化，如图 8 所示。

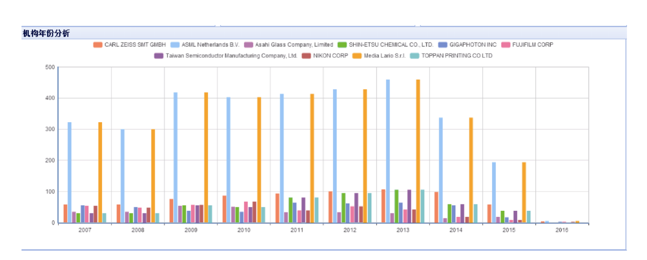


图 8　极紫外光刻技术主要专利申请人随时间变化趋势分析示例

极紫外光刻技术专利预警平台的建设过程表明："专题专利预警分析平台"可以定制化实现专题的专利数据采集、分类、跟踪、分析，并提供了优良的预警分析数据结果。

5　专题专利预警平台的后续工作

专题专利预警平台建设方案为需要进行专题长期跟踪预警分析、专题数据再利用的专题研究、专题管理、专题情报等工作提供一种可行的定制化解决途径。本文针对专题专利预警平台的建设方案、技术实现进行了阐述，并基于"极紫外光刻专利预警分析平台"的建设过程进行实验分析，验证了专题专利预警平台建设方案的可行性和有效性。基于这一思路开发的"专题专利预警平台"还存在需要改进和提高的地方，如：基础平台的优化、数据处理的全自动化、数据分析的指标化、内容挖掘的关联实现等，这也是未来在实践中的工作重点。

参考文献：

[1] 张勇. 专利预警分析——从管控风险到决胜创新[M]. 第 1 版. 北京：知识产权出版社，2015: 26-28. (Zhang Yong. Patent Pre-Waring——from Risk Management to Innovative Competition [M]. The 1st Edition. Beijing: Intellectual Property Publishing House, 2015: 26-28.)

[2] 张智雄，张晓林，刘建华，等. 网络科技信息结构化监测的思路和技术方法实现[J]. 中国图书馆学报，2014, 40(4): 4-15. (Zhang Zhixiong, Zhang Xiaolin, Liu Jianhua, et al. The Ideas and Methods of Structural Monitoring of the Scientific and Technical Information Resources on the Web[J]. Journal of Library Science in China, 2014, 40(4): 4-15.)

[3] 祝忠明，马建霞，常宁，等. 基于 DSpace 构建学科知识库系统的研究与实践[J]. 现代图书情报技术, 2006 (7): 10-14. (Zhu Zhongming, Ma Jianxia, Chang Ning, et al. An Implementation of a DSpace-based Disciplinary Repository System [J]. New Technology of Library and Information Service, 2006 (7): 10-14.)

[4] DuraSpace. DSpace 4.x Documentation [R/OL]. [2014-08-20]. http://www.yok.gov.tr/documents/7166509/7180015/DSpace-Manual+4.x.pdf/4ac490ee-9a24-4edd-90b7-a894134c9641.

[5] 王丽. 开源/免费工具比较及专利分析全流程解决方案研究[J]. 情报理论与实践，2016, 39(1): 118-122. (Wang Li. Solution Research of Open-Source/Fee-Free Tools Used in Full-Process Patent Analysis [J]. Information Studies: Theory & Application, 2016, 39(1): 118-122.)

[6] Verborgh R, De Wilde M. Using OpenRefine[M]. Packt Open Source, 2013: 21-64.

[7] ECharts 团队. ECharts 参考手册 [R/OL]. [2016-05-20]. http://echarts.baidu.com/echarts2/doc/doc.html. (ECharts Team. Getting Started [R/OL]. [2016-05-20]. http://echarts.baidu.com/echarts2/doc/doc.html.)

[8] Centre for Science and Technology Studies, Leiden University. VOSviewer-Getting Started [OL]. [2015-10-10]. http://www.vosviewer.com/getting-started.

作者贡献声明：

王丽：提出研究思路，设计研究方案、技术实现方案，进行实验，采集、清洗和分析数据，论文起草、修改及最终版本修订；
丁迎杰：技术实现与优化，论文修改；
吴鸣：提出研究思路。

利益冲突声明：

所有作者声明不存在利益冲突关系。

支撑数据：

支撑数据由作者自存储, E-mail: wangli@mail.las.ac.cn。
[1] 王丽. CAScode.zip. 专题专利预警平台原创代码.
[3] 王丽. Dspace_import.zip. 极紫外光刻技术专利预警平台数据.

收稿日期: 2016-06-29
收修改稿日期: 2016-07-28

Building Subject-Based Early Warning System for Patents

Wang Li[1, 2] Ding Yingjie[1] Wu Ming[1]
[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)
[2] (University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: [**Objective**] This paper proposes a Subject-Based Early Warning System for Patents, which provides a solution to long-term project tracking, early warning analysis, and data reuse. [**Methods**] Subject-Based Early Warning System for Patents integrated some open source systems and tools (e.g.: DSpace, OpenRefine, ECharts, VOSviewer, etc.) and developed the functions of data storing, tracking, classifying, cleansing, analyzing and managing. [**Results**] First, we constructed the new system with the subject of extreme ultraviolet lithography. Second, we examined the feasibility and effectiveness of the new system. [**Limitations**] The data processing automation, data analysis indicators, and content mining need to be optimized. [**Conclusions**] The proposed system could track, manage and utilize patent information effectively.

Keywords: Subject-based system Patent early warning System constraction

第 29 届全国计算机信息管理学术研讨会在甘肃省张掖召开

中国科学技术情报学会信息技术专业委员会于 2016 年 10 月 8 日–10 日在甘肃省张掖市成功举办了“第二十九届全国计算机信息管理学术研讨会”。此次会议由北京超选智能科技研究院承办，来自全国科研院所、高校和企业的 40 余名代表参加了会议。本次会议沿袭专委会 30 多年的学术探索和技术引领传统，以“推进科技信息服务技术的创业创新”为主题，开展了广泛、深入的交流讨论。会议分为专题报告和论文交流两个环节，来自中国科学技术信息研究所、中国国防科技信息中心、航空工业发展研究中心、解放军南京政治学院、北京大学、北京师范大学等研究机构和万方数据股份有限公司、北京超选智能科技研究院等企业的专家学者，就科技情报领域最新热点和发展趋势进行了深入探讨。研讨内容涵盖了现代信息服务技术发展、知识服务平台建设、知识资源服务中心建设、科技信息行业的大数据工程应用、智库建设、科技信息行业资源融合、深度学习和人工智能技术在信息行业的探索实践。在研讨过程中，与会专家就科技信息新型技术应用、科技信息服务未来发展方向、科技信息人才队伍建设等话题进行了深入交流并提出了宝贵建议。本届会议共选用优秀论文 26 篇，反映了近年我国科技信息领域所取得的研究成果，具有较高的学术水平；能够反映当前我国科技信息界信息技术的应用动向和最新进展。会议的成功召开，进一步推进了科技信息技术体系的基础性研究，积极探索了科技信息技术的应用与创新。会议充分体现了随着国家创新驱动发展战略的深入实施，科技信息行业必将继续发扬信息技术应用前沿的传统，在大力推进大众创业、万众创新的新时期继续发挥重要作用。

(本刊讯)